Year: 2014

On a class of high order schemes for hyperbolic problems

Abgrall, Rémi

Abstract: This paper provides a review about a family of non oscillatory and parameter free finite element type methods for advection-diffusion problems. Due to space limitation, only the scalar hyperbolic problem is considered. We also show that this class of schemes can be interpreted as finite volume schemes with multidimensional fluxes.

Originally published at:
Abgrall, Rémi (2014). On a class of high order schemes for hyperbolic problems. In: Jang, Sun Young; Kim, Young Rock; Lee, Dae-Woong; Yie, Ikkwon. Proceedings of the International Congress of Mathematicians, Seoul 2014. Seoul: Korea Science and Technology Center, 699-725.

SEOUL ICM 2014
INTERNATIONAL CONGRESS OF MATHEMATICIANS

Proceedings of the International Congress of Mathematicians

Seoul 2014

VOLUME IV
Invited Lectures

Editors

Sun Young Jang
Young Rock Kim
Dae-Woong Lee
Ikkwon Yie

KM KYUNG MOON SA

Editors

Sun Young Jang, University of Ulsan
Young Rock Kim, Hankuk University of Foreign Studies
Dae-Woong Lee, Chonbuk National University
Ikkwon Yie, Inha University

Technical Editors

Young Rock Kim, The Korean TEX Society
Hyun Woo Kwon, The Korean TEX Society

Proceedings of the International Congress of Mathematicians
August 13–21, 2014, Seoul, Korea

Published by

KYUNG MOON SA Co. Ltd.
174, Wausan-ro Mapo-gu Seoul, Korea
Tel: +82-2-332-2004 Fax: +82-2-336-5193
E-mail: kyungmoon@kyungmoon.com
Homepage: www.kyungmoon.com

ISBN 978-89-6105-807-0
ISBN 978-89-6105-803-2 (set)

Printed in Korea

Contents

12. Probability and Statistics

13. Combinatorics

17. Mathematics in Science and Technology

18. Mathematics Education and Popularization of Mathematics

19. History of Mathematics

On a class of high order schemes for hyperbolic problems

Rémi Abgrall

Abstract. This paper provides a review about a family of non oscillatory and parameter free finite element type methods for advection-diffusion problems. Due to space limitation, only the scalar hyperbolic problem is considered. We also show that this class of schemes can be interpreted as finite volume schemes with multidimensional fluxes.

Mathematics Subject Classification (2010). 65, 76.

Keywords. Numerical approximation of hyperbolic problems, Non oscillatory schemes, Unstructured meshes, High order methods

1. Introduction

We are interested in the numerical solution of parabolic type equations in which the elliptic terms play an important role only at some locations of the computational domain. To make things more precise, our target are the Navier-Stokes equations in the compressible regime. These systems of partial differential equations are supplemented by initial and boundary conditions. In particular, at solid walls, the velocity is set to zero and the temperature behavior is specified. Thus depending on the Reynolds number, the viscous terms have an effect that is sensitive on a more or large range. Far enough from the walls, where the viscous effects are less prominent, it is mainly the hyperbolic part that plays the major role, and thus, depending on the flow conditions, thin zone with very steep gradients may exist with a shock-like structure.

Our goal is to approximate the solution every where, with a parameter free method, so that the solution is oscillation free, with a uniform accuracy. In addition, we want to handle complicated geometries, so that the method use unstructured meshes.

How can this program be achieved? In the following, we focus on steady problems, and to make things simpler, we focus on the scalar problem:

$$\text{div } \mathbf{f}(u) = 0 \tag{1a}$$

subjected to

$$\min(\nabla_u f(u) \cdot \mathbf{n}(\mathbf{x}), 0)(u - g) = 0 \text{ on } \partial\Omega \tag{1b}$$

In (1b), $\mathbf{n}(\mathbf{x})$ is the outward unit vector at $\mathbf{x} \in \partial\Omega$ (thus we assume enough regularity for Ω). The case of the advection-diffusion problem

$$\text{div } \mathbf{f}(u) - \text{div}(\mathbf{K}\nabla u) = 0 \tag{2a}$$

Proceedings of the International Congress of Mathematicians, Seoul, 2014

subjected to boundary condition of the Dirichlet type

$$u = g \text{ on } \Gamma_D \tag{2b}$$

and Neuman-like conditions

$$(\mathbf{K}\nabla u)\cdot \mathbf{n}(\mathbf{x}) = h(\mathbf{x}) \text{ on } \Gamma_N \tag{2c}$$

is done in a similar way except for some technicalities about the diffusion term, see [4]. Extensions to the system case can be found in [5] for the pure hyperbolic case and [3] for the Navier Stokes equations.

Here the notations are standard: g and h are regular enough functions, Γ_D and Γ_N are non overlapping regular enough subsets of $\partial\Omega$, and $\Gamma_D \cup \Gamma_N = \partial\Omega$. From now on, we assume that Ω has a polyhedric boundary, and more over $\Omega_h = \Omega$ for the chosen family of triangulations in order to simplify. These assumptions are by no mean essential. We denote by $\mathcal{E}_h$ the set of edges/faces of $\mathcal{T}_h$ that are contained in $\partial\Omega$, and $\mathcal{K}$ stands either for an element K or a face/edge e.

In the finite element setting, there exists several variational formulations of this class of problems. The classical ones can be defined in three steps. We are given a family of meshes denoted by $(\mathcal{T}_h)_{h\in\mathcal{H}}$. These meshes are made of elements denoted generically by K. The parameter h, as usual, denotes the maximum of the diameters of K, $K \in \mathcal{T}_h$. The meshes can be geometrically conformal or not. Then we need to define the trial function space, denoted by U_h and a test function V_h. The last step is to define a bi-linear form a on $U_h \times V_h$, as well as form ℓ defined on V_h. As usual, we assume that the spaces U_h and V_h encode some of the boundary conditions, while the others are encoded in ℓ. The problem is to find $u_h \in U_h$ such that a for any $v_h \in V_h$, we have

$$a(u_h, v_h) = \ell(v_h).$$

A first example example is given by the streamline diffusion method [12, 13] for which there are two possible interpretations. In the first one, we consider a Petrov Galerkin formulation, .i.e we take

$$U_h = \{u_h \in H^1(\Omega) \text{ such that for any } K \in \mathcal{T}_h, \quad u_h\big|_K \in \mathbb{P}^r(K)\} \cap C^0(\overline{\Omega})$$

and

$$\begin{aligned} V_h = &\{v_h \in L^2(\Omega), \text{ such that for any } K \in \mathcal{T}_h, \text{ there exists } w_h \in U_h, \\ &v_h = w_h + h_K \tau_K \nabla_u \mathbf{f}(u_h)\nabla w_h\} \end{aligned}$$

and

$$a_{\text{SUPG1}}(u_h, v_h) = \int_\Omega v_h \ \text{div}\, \mathbf{f}(u_h) + \sum_{e\in\mathcal{E}_h} \int_e v_h\big(\hat{\mathbf{f}}_{\mathbf{n}}(g, u_h) - \mathbf{f}(u_h)\cdot\mathbf{n}\big) = \int_\Omega f v_h. \tag{3a}$$

Here, $\hat{\mathbf{f}}$ is a consistent upwind numerical flux. The second interpretation is to take $V_h = U_h$ and use, instead of a_{SUPG1} the form a_{SUP2} defined by

$$a_{\text{SUPG2}}(u_h, v_h) = \int_\Omega v_h \ \text{div}\, \mathbf{f}(u_h) + \sum_K h_K \int_K \big(\nabla_u f(u_h)\nabla v_h\big)\tau_K\big(\nabla_u f(u_h)\nabla u_h\big)$$

$$+\sum_{e\in\mathcal{E}_h}\int_e v_h\big(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h)-\mathbf{f}(u_h)\cdot\mathbf{n}\big). \tag{3b}$$

This can be as a Galerkin approximation of a modified equation, namely

$$\text{div}\,\mathbf{f}(u)-\text{div}\,\Big(h\nabla_u\mathbf{f}(u)\otimes(\tau\nabla_u\mathbf{f}(u))\nabla u\Big)=0. \tag{3c}$$

In (3), the parameters τ_K are positive functions (typically constant per element) and in (3c) the function τ and h are defined by their restrictions on each element.

We can play further with the trial and test spaces. If one removes the continuity assumption, then we have a Discontinuous Galerkin formulation, i.e. $U_h=V_h$ with

$$U_h=\{u_h\in L^2(\Omega),\ \text{such that for any } K\in\mathcal{T}_h,\quad u_h\big|_K\in\mathbb{P}^r(K)\}$$

and

$$a(u_h,v_h)=\sum_{K\in\mathcal{T}_h}\Big(-\int_K\nabla v_h\cdot\mathbf{f}(u^h)+\int_{\partial K}v_h\hat{\mathbf{f}}_{\mathbf{n}}\big((u_h)_{|K},(u_h)_{|K^-}\big)\Big) \tag{4a}$$

where K^- denotes generically the element(s) that are on the other side of the faces of ∂K. Another formulation is

$$\begin{aligned}a(u_h,v_h)=&\sum_{K\in\mathcal{T}_h}\Big(-\int_K\nabla v_h\cdot\mathbf{f}(u^h)+\int_{\partial K}v_h\hat{\mathbf{f}}_{\mathbf{n}}\big((u_h)_{|K},(u_h)_{|K^-}\big)\Big)\\&+\sum_K h_K\int_K\big(\nabla_u f(u_h)\nabla v_h\big)\tau_K\big(\nabla_u f(u_h)\nabla u_h\big)\end{aligned} \tag{4b}$$

In (4), the Dirichlet boundary conditions are set weakly, as in (3), by setting $u_h=g$ on the parts of ∂K which belongs to inflow part of $\partial\Omega$.

The space U_h and V_h can be independently chosen, as well as a and ℓ, provided the variational problem is consistent with the problem (1), and of course the numerical method is stable. Formal accuracy is obtained via the choice the polynomial degree r, and effective accuracy is related to the stability of the scheme in suitable norm. Hence a natural question is: can we define U_h, V_h and the forms a and ℓ such that in addition with consistency and accuracy, we can also have non oscillatory properties. In the case of the streamline methods, this last property is obtained by modifying the formulation by adding a dissipation operator which is parameter dependent. In the case of the Discontinuous Galerkin method, this property is obtained via a proper choice of the arguments in $\hat{\mathbf{f}}_{\mathbf{n}}$, see [7, 8]. We note that only the averages in K are controlled. In both cases this is obtained by introducing some genuine non linearity in the scheme, i.e. even if (1) is a linear problem, the scheme will be non linear.

In this paper, we show that, by introducing a solution-dependent operator χ from $U_h\cap C^0(\Omega)$ to $L^2(\Omega)$, the variational problem with a defined by

$$a(u_h,v_h)=\sum_K\int_K\chi_u^h(v_h)\text{div}\,\mathbf{f}(u^h)+\sum_{e\in\mathcal{E}}\int_e v_h(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h)-\mathbf{f}(u_h)\cdot\mathbf{n}) \tag{5}$$

enables to get all the properties. The rest of this paper is organized as follow: inspired by a rewriting of (3), we introduce the residual distribution schemes. We provide a simple criteria which guaranty a Lax-Wendroff type theorem, provide a simple criteria that guaranties formal accuracy, show how the choice of norms guaranty the effective accuracy, and provide several examples of schemes. In the last part, we show how these schemes can also be interpreted as finite volume schemes and we provide explicit formula.

2. Formulation of residual distribution schemes

These schemes have been introduced by P.L. Roe in [17] in one dimension, and [18] in the multidimensional case. As we see, there are many common points with the streamline method, the difference is that we try to combine ideas from the finite element community and from the finite volume one. The first scheme of this kind was probably designed by R. Ni [16] where introduce a particular version of the Lax Wendroff scheme.

2.1. Definition, connection to finite element methods. We make the standard remark that, for any internal degree of freedom σ, if φ_σ is the Lagrange basis function associated to σ, (3b) can be written as:

$$a_{\text{SUPG2}}(u_h,\varphi_\sigma)=\sum_K\Big(\int_K\varphi_\sigma\nabla\cdot\mathbf{f}(u_h)+h_K\int_K\big(\nabla_u\mathbf{f}(u_h)\nabla\varphi_\sigma\big)\tau_K\big(\nabla_u\mathbf{f}(u_h)\nabla u_h\big)\Big)\\+\sum_{e\in\mathcal{E}_h}\int_e\varphi_\sigma(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h)-\mathbf{f}(u_h)\cdot\mathbf{n}).$$

Since the support of φ_σ is made of all the elements K that share σ, we have for any degree of freedom σ:

$$a_{SUPG2}(u_h,\varphi_\sigma)\\=\sum_{K\ni\sigma}\Big(\int_K\varphi_\sigma\nabla\cdot\mathbf{f}(u_h)+h_K\int_K\big(\nabla_u\mathbf{f}(u_h)\nabla\varphi_\sigma\big)\tau_K\big(\nabla_u\mathbf{f}(u_h)\nabla u_h\big)\Big)\\\sum_{e\in\mathcal{E}_h,\sigma\in e}\int_e\varphi_\sigma(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h)-\mathbf{f}(u_h)\cdot\mathbf{n})$$

and notice that

1. for any K,
$$\sum_{\sigma\in K}\Big(\int_K\varphi_\sigma\nabla\cdot\mathbf{f}(u_h)+h_K\int_K\big(\nabla_u\mathbf{f}(u_h)\nabla\varphi_\sigma\big)\tau_K\big(\nabla_u\mathbf{f}(u_h)\nabla u_h\big)\Big)=\int_{\partial K}\mathbf{f}(u_h)\cdot\mathbf{n},$$
2. for any $e\in\mathcal{E}_h$,
$$\sum_{\sigma\in e}\int_e\varphi_\sigma(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h)-\mathbf{f}(u_h)\cdot\mathbf{n})=\int_e(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h)-\mathbf{f}(u_h)\cdot\mathbf{n}).$$

This is true because $\sum_{\sigma\in K}\varphi_\sigma(\mathbf{x})=1$ and thus $\sum_{\sigma\in K}\nabla\varphi_\sigma(\mathbf{x})=0$ for all $x\in\mathcal{K}$.

We define the total residual for element and edges the quantities:

$$\Phi^K:=\int_{\partial K}\mathbf{f}(u_h)\cdot\mathbf{n},\text{ and }\Phi^e:=\int_e(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h)-\mathbf{f}(u_h)\cdot\mathbf{n}).\tag{6}$$

A residual distribution scheme is defined by the sub-residuals that are "sent" to the degrees of freedom σ by an element K (resp. a boundary edge e). We denote them by $\Phi_\sigma^K(u^h_{|K})$ (resp. $\Phi_\sigma^e(u^h_{|e})$). The scheme writes, for any internal degree of freedom σ,

$$\sum_{K\ni\sigma} \Phi_\sigma^K(u^h_{|K}) = 0, \tag{7a}$$

and for any degree of freedom on the boundary,

$$\sum_{K\ni\sigma} \Phi_\sigma^K(u^h_{|K}) + \sum_{e\ni\sigma} \Phi_\sigma^e(u^h_{|e}) = 0. \tag{7b}$$

We assume that the following structure condition holds true:

$$\forall \sigma\in K, \qquad \sum_{\sigma\in K} \Phi_\sigma^K(u^h_{|K}) = \int_{\partial K} \mathbf{f}(u_h)\cdot\mathbf{n} \qquad (=\Phi^K), \tag{8a}$$

$$\forall e\in \mathcal{E}_h, \qquad \sum_{\sigma\in e} \Phi_\sigma^e(u^h_{|K}) = \int_e (\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h) - \mathbf{f}(u_h)\cdot\mathbf{n}). \qquad (=\Phi^e) \tag{8b}$$

We see that the SUPG method is a particular case of such scheme. There is a lot of freedom in defining the sub-residuals $\Phi_\sigma^K(u^h_{|K})$ and $\Phi_\sigma^e(u^h_e)$, we will show how we can take advantage of this freedom to achieve our goal. Note that in the definition of the sub-residual, we have implicitly assumed that only the degrees of freedom with K or e are necessary to define these quantities: the stencil of the method is the most possible compact which is a good point for the parallelization of the method.

Another example of sub-residual are the Galerkin residuals defined by: on the element K

$$\Phi_\sigma^{G,K} = \int_K \varphi_\sigma \mathrm{div}\,\mathbf{f}(u^h = -\int_K \nabla\varphi_\sigma\cdot\mathbf{f}(u^h) + \int_{\partial K} \varphi_\sigma \mathbf{f}(u^h)\cdot\mathbf{n}, \tag{9a}$$

and on the boundary face e:

$$\Phi_\sigma^{G,e} = \int_e \varphi_\sigma(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h) - \mathbf{f}(u_h)\cdot\mathbf{n}) \tag{9b}$$

We see that both $\{\Phi_\sigma^{G,K}\}_{\sigma\in K}$ and $\{\Phi_\sigma^{G,e}\}_{\sigma\in e}$ satisfy (8) with the same value of the total residual. Unfortunately, the scheme (7) with the Galerkin residual is widely unstable,

2.2. Structure conditions. For any w^h (not necessarily a solution of (7) if it exists), and any test function v^h, we have (setting $v^h_\sigma = v^h(\sigma)$):

$$\begin{aligned} S &:= \sum_{\sigma\notin\partial\Omega} v^h_\sigma\left(\sum_{K\ni\sigma}\Phi_\sigma^K(w^h_{|K})\right) + \sum_{\sigma\in\partial\Omega} v^h_\sigma\left(\sum_{K\ni\sigma}\Phi_\sigma^K(w^h_{|K}) + \sum_{e\ni\sigma, e\in\mathcal{E}_h}\Phi_\sigma^e(w^h_{|e})\right) \\ &= \sum_K\left(\sum_{\sigma\in K} v^h_\sigma\Phi_\sigma^K(w^h_{|K})\right) + \sum_{e\in\mathcal{E}_h}\left(\sum_{\sigma\in e} v^h_\sigma\Phi_\sigma^e(w^h_{|K})\right) \\ &= -\int_\Omega v^h\nabla\cdot\mathbf{f}(u^h) + \int_{\partial\Omega} v^h\hat{\mathbf{f}}_{\mathbf{n}}(g,w^h) \end{aligned} \tag{10}$$

$$+\sum_K \sum_{\sigma\in K} v^h_\sigma\big(\Phi^K_\sigma(w^h_{|K}) - \Phi^{G,K}_\sigma(w^h_{|K})\big)$$
$$+\sum_{e,e\in\mathcal{E}_h} \sum_{\sigma\in e} v^h_\sigma\big(\Phi^e_\sigma(w^h_{|K}) - \Phi^{G,e}_\sigma(w^h_{|K})\big)$$

thanks to (9). Then, since (recall $\mathcal{K}$ represents either a generic element or a generic member of $\mathcal{E}_h$)

$$\sum_{\sigma\in\mathcal{K}}\left(\Phi^{\mathcal{K}}_\sigma(w^h_{|\mathcal{K}}) - \Phi^{G,\mathcal{K}}_\sigma(w^h_{|K})\right) = 0,$$

(10) becomes, denoting by n_K and n_e the number of degree of freedom in K and e:

$$\begin{aligned} S = &-\int_\Omega \nabla v^h \cdot \mathbf{f}(u^h) + \int_\Omega v^h \hat{\mathbf{f}}_{\mathbf{n}}(g, w^h)\\ &+\sum_K \frac{1}{n_K!}\sum_{\sigma,\sigma'\in K}\big(v^h_\sigma - v^h_{\sigma'}\big)\big(\Phi^K_\sigma(w^h_{|K}) - \Phi^{G,K}_\sigma(w^h_{|K})\big)\\ &+\sum_{e\in\mathcal{E}_h}\frac{1}{n_e!}\sum_{\sigma,\sigma'\in e}\big(v^h_\sigma - v^h_{\sigma'}\big)\big(\Phi^e_\sigma(w^h_{|e}) - \Phi^{G,e}_\sigma(w^h_{|e})\big) \end{aligned} \tag{11}$$

This relation is fundamental in our analysis.

2.2.1. Conservation. In [6], we prove the following result:

Theorem 2.1. *Assume the family of meshes $\mathcal{T} = (\mathcal{T}_h)_{h\in\mathcal{H}}$ is regular. We assume that the residuals $\{\Phi^{\mathcal{K}}_\sigma\}_{\sigma\in\mathcal{K}}$, for $\mathcal{K}$ an element or a boundary element of $\mathcal{T}_h$, satisfy:*

- *For any $M \in \mathbb{R}^+$, there exists a constant C which depends only on the family of meshes $\mathcal{T}_h$ and M such that for any $u_h \in U_h$ with $||u^h||_\infty \leq M$, then*

$$\big\|\Phi^{\mathcal{K}}_\sigma({u^h}_{|\mathcal{K}})\big\| \leq C \sum_{\sigma,\sigma'\in\mathcal{K}} |u^h_\sigma - u^h_{\sigma'}|$$

- *they satisfy the conservation property* (8).

Then if there exists a constant C_{max} such that the solutions of the scheme (7) *satisfy $||u^h||_\infty \leq C_{max}$ and a function $v \in L^2(\Omega)$ such that $(u^h)_h$ or at least a sub-sequence converges to v in $L^2(\Omega)$, then v is a weak solution of* (1)

Proof. The proof can be found in [6], it uses (11) and some adaptation of the ideas of [14]. □

We can also state condition for entropy inequalities:

Proposition 2.2. *Let $(U, \mathbf{G})$ be an couple entropy-flux for* (1) *and $\hat{\mathbf{G}}_{\mathbf{n}}$ an upwind numerical entropy flux consistent with $\mathbf{G}\cdot\mathbf{n}$. Assume that the residuals satisfy: for any element K,*

$$\sum_{\sigma\in K} U(u_\sigma)\Phi^K_\sigma \leq \int_{\partial K} \mathbf{G}(u^h_{|\mathcal{K}})\cdot\mathbf{n} \tag{12a}$$

and for any boundary edge e,

$$\sum_{\sigma\in e} U(u_\sigma)\Phi_\sigma^e \leq \int_e \big(\hat{\mathbf{G}_{\mathbf{n}}}(u^h_{|e}, g) - \mathbf{G}(u^h_{|K})\cdot \mathbf{n}\big). \tag{12b}$$

Then, under the assumptions of the theorem 2.1, the limit weak solution also satisfies the following entropy inequality: for any $\varphi \in C^1(\Omega)$, $\varphi \geq 0$,

$$-\int_\Omega \nabla\varphi \cdot \mathbf{G}(u) + \int_{\partial\Omega} \hat{\mathbf{G}}_{\mathbf{n}}(u, g) \leq 0.$$

Proof. The proof is similar to that of theorem 2.1. □

2.2.2. Accuracy. In most cases, assuming a smooth solution of (1), the formal accuracy analysis is done by checking how large is the error made when plugging the exact solution into the scheme. This is carried out using Taylor expansions, and the geometry of the computational stencil plays an important role. When the mesh has no particular symmetry, this leads to nowhere. Instead of looking to how far the numerical scheme departs from the strong form of the PDE, it is much more flexible to look at how for it departs its weak form, i.e. instead of checking $\text{div}\,\mathbf{f}(u) = 0$, it is better to test, for any φ smooth enough, $\int_\Omega \varphi\, \text{div}\,\mathbf{f}(u) = 0$, of course after using the Green formula.

In practice, we define the truncation error

$$\mathcal{E}(w^h, v^h) = \sum_{\sigma\notin\partial\Omega} v^h_\sigma \left(\sum_{K\ni\sigma} \Phi_\sigma^K(w^h_{|K})\right),$$

and consider

$$\mathcal{E}(w^h) = \max_{v^h\in V^h, ||v^h||_{W^{1,\infty}}=1} \mathcal{E}(w^h, v^h). \tag{13}$$

We can then extend the classical definition of accuracy:

Definition 2.3 (Accuracy). We say that the scheme (7) is $r+1$-th order accurate if, for any smooth solution $u_{ex} \in C^{r+1}(\overline{\Omega})$ of (1), $\mathcal{E}(u^h_{ex}) \leq C\, h^{r+1}$. The constant C only depend on the family $\mathcal{T}$, the regularity of $\mathbf{f}$, on the $r+1$ derivative of u, and the boundary conditions.

Using (11), we see that, for any v^h

$$\mathcal{E}(u^h_{ex}, v^h) = -\int_\Omega \nabla v^h \cdot \mathbf{f}(u^h_{ex}) + \int_\Omega v^h \hat{\mathbf{f}}_{\mathbf{n}}(g, u^h_{ex}) \tag{14}$$

$$+\sum_K \frac{1}{n_K!} \sum_{\sigma,\sigma'\in K} \big(v^h_\sigma - v^h_{\sigma'}\big)\big(\Phi_\sigma^K((u^h_{ex})_{|K}) - \Phi_\sigma^{G,K}((u^h_{ex})_{|K})\big) \tag{15}$$

$$+\sum_{e\in\mathcal{E}_h} \frac{1}{n_e!} \sum_{\sigma,\sigma'\in e} \big(v^h_\sigma - v^h_{\sigma'}\big)\big(\Phi_\sigma^e((u^h_{ex})_{|K}) - \Phi_\sigma^{G,e}((u^h_{ex})_{|e})\big) \tag{16}$$

For the *steady* problem (1), we have the following result:

Lemma 2.4. *Let us recall that $\Omega \subset \mathbb{R}^d$ and is bounded.*

If the solution u_{ex} of the steady problem (1) *is C^{r+1}, then*

(1) $\Phi_\sigma^{G,K}((u_{ex}^h)_{|K}) = O(h^{r+d})$,

(2) $\Phi_\sigma^{G,e}((u_{ex}^h)_{|e}) = O(h^{r+d-1})$

(3) *if the numerical flux* $\hat{\mathbf{f}}$ *is Lipschitz,* $-\int_\Omega \nabla v^h \cdot \mathbf{f}(u_{ex}^h) + \int_\Omega v^h \hat{\mathbf{f}}_{\mathbf{n}}(g, u_{ex}^h) = O(h^{r+1})$,

Proof. We start by showing the first result. The proof of the second one is similar and is omitted.

Since $u_{ex} \in C^{r+1}$, we have $\text{div}\, \mathbf{f}(u_{ex}) = 0$ in a strong sense, thus for any $K \in \mathcal{T}_h$ and any σ,

$$\int_K \varphi_\sigma \text{ div } \mathbf{f}(u_{ex}) = -\int_K \nabla\phi_\sigma \cdot \mathbf{f}(u_{ex}) + \int_{\partial K} \phi_\sigma \mathbf{f}(u_{ex}) \cdot \mathbf{n} = 0.$$

We can subtract this relation to $\Phi_\sigma^{G,K}(u_{ex}^h)$ and get:

$$\Phi_\sigma^{G,K}(u_{ex}^h) = -\int_K \nabla\varphi_\sigma \cdot \Big(\mathbf{f}(u_{ex}^h) - \mathbf{f}(u_e)\Big) + \int_{\partial K} \varphi_\sigma \Big(\mathbf{f}(u_{ex}^h) - \mathbf{f}(u_e)\Big).$$

Since the mesh is regular, we have:

$$|K| = O(h^d), \qquad \nabla\varphi_\sigma = O(h^{-1}), \qquad |\partial K| = O(h^{d-1})$$

and since the flux $\mathbf{f}$ is C^1, we have

$$\mathbf{f}(u_{ex}^h) - \mathbf{f}(u_e) = O(h^{k+1}).$$

Gathering the pieces together, we get:

$$\Big|\Phi_\sigma^{G,K}(u_{ex}^h)\Big| \leq C\Big(h^d \times h^{-1} \times h^{k+1} + h^{d-1} \times 1 \times h^{k+1}\Big) = O(h^{k+d}).$$

The third inequality is obtained in a similar manner: From (1), we have for any v^h, setting $\Gamma^- = \{\mathbf{x} \in \partial\Omega, \nabla_u \mathbf{f}(u) \cdot \mathbf{n} < 0\}$,

$$-\int_\Omega \nabla v^h \cdot \mathbf{f}(u_{ex}) + \int_{\Gamma^-} v^h \mathbf{f}(u_{ex}) \cdot \mathbf{n} = 0$$

so that

$$\begin{aligned} -\int_\Omega \nabla v^h \cdot \mathbf{f}(u_{ex}^h) + \int_\Omega v^h \hat{\mathbf{f}}_{\mathbf{n}}(g, u_{ex}^h) \\ &= -\int_\Omega \nabla v^h \cdot \big(\mathbf{f}(u_{ex}^h) - \mathbf{f}(u_{ex})\big) + \int_{\partial\Omega} v^h \Big(\hat{\mathbf{f}}_{\mathbf{n}}(g, u_{ex}^h) - \mathbf{f}(u_{ex}^h) \cdot \mathbf{n}\Big) \\ &= (I) + (II) \end{aligned}$$

Using again the same arguments, since the numerical flux is Lipschitz continuous, we see that both (I) and (II) are of the order of $O(h^{k+1}) \times ||v^h||_{W^{1,\infty}(\Omega)}$. □

Then, we have:

Proposition 2.5. *Under the assumptions of Lemma 2.4 and assuming that the family of meshes* $\mathcal{T}_h$ *is regular, the residuals satisfy:*

$$\textit{for all } \sigma \textit{ and all } \mathcal{K} = K \textit{ or } e, \Phi_\sigma^{\mathcal{K}}((u_{ex})_{|\mathcal{K}}) = O(h^{r+D}) \tag{17}$$

where $D = d$ *for elements and* $D = d - 1$ *for* $e \in \mathcal{E}$*, then the scheme is formally* $r + 1$ *accurate.*

Proof. $\mathcal{E}(u^h_{ex}, v^h)$ is the sum of

$$-\int_\Omega \nabla v^h \cdot \mathbf{f}(u^h_{ex}) + \int_\Omega v^h \hat{\mathbf{f}}_{\mathbf{n}}(g, u^h_{ex})$$

which is $O(h^{r+1})$ by lemma 2.4 and

$$\sum_K \frac{1}{n_K!} \sum_{\sigma,\sigma' \in K} \big(v^h_\sigma - v^h_{\sigma'}\big)\big(\Phi^K_\sigma(w^h_{|K}) - \Phi^{G,K}_\sigma(w^h_{|K})\big) \\ + \sum_{e\subset\Omega} \frac{1}{n_e!} \sum_{\sigma,\sigma'\in e} \big(v^h_\sigma - v^h_{\sigma'}\big)\big(\Phi^e_\sigma(w^h_{|K}) - \Phi^{G,e}_\sigma(w^h_{|K})\big)$$

Since the mesh is regular, the number of elements in the mesh is $O(h^{-d})$ and the number of boundary elements is $O(h^{d-1})$. Since $v \in W^{1,\infty}$, its Lagrange interpolant satisfy

$$\big|v^h_\sigma - v^h_{\sigma'}\big| \leq h||v^h||_{W^{1,\infty}}$$

and $\sup_h ||v^h||_{W^{1,\infty}}$ is bounded by a constant that depends on $\mathcal{T}$ and $||v||_{1,\infty}$. Then we see that

$$\begin{aligned}\Big|\sum_K \frac{1}{n_K!} \sum_{\sigma,\sigma' \in K} \big(v^h_\sigma - v^h_{\sigma'}\big)\big(\Phi^K_\sigma(w^h_{|K}) - \Phi^{G,K}_\sigma(w^h_{|K})\big) \\ + \sum_{e\subset\partial\Omega} \frac{1}{N_e!} \sum_{\sigma,\sigma'\in e} \big(v^h_\sigma - v^h_{\sigma'}\big)\big(\Phi^e_\sigma(w^h_{|K}) - \Phi^{G,e}_\sigma(w^h_{|K})\big)\Big| \\ \leq C\big(h^{-d} \times h \times h^{d+r} + h^{-d+1} \times h \times h^{r+d-1}\big) \\ \leq Ch^{r+1}\end{aligned}$$

□

3. Construction of monotonicity preserving arbitrary accurate schemes

We start by a basic remark that goes at least back to A. Harten [11], and we rephrase it in the Residual Distribution framework.

Lemma 3.1. *Assume that the residual (for element and edges) write, for any degree of freedom,*

$$\Phi^{\mathcal{K}}_\sigma(u_h) = \sum_{\sigma'\ni\mathcal{K}} c^K_{\sigma\sigma'}(u_\sigma - u_{\sigma'}), \tag{18}$$

then the iterative scheme

$$u^{n+1}_\sigma = u^n_\sigma - \omega_\sigma\Big(\sum_{K\ni\sigma} \Phi^K_\sigma + \sum_{e\ni\sigma} \Phi^e_\sigma\big)$$

admits a local maximum principle if

- *for any σ, σ', $c^K_{\sigma\sigma'} \geq 0$,*
- $\omega_\sigma \left(\sum_{K\ni\sigma}\sum_{\sigma'\in K} c^K_{\sigma\sigma'} + \sum_{\sigma'\in K} c_{\sigma\sigma'}\right) \leq 1$

Proof. It is clear that:

$$\sum_{K\ni\sigma}\Phi_\sigma^K+\sum_{e\ni\sigma, e\in\mathcal{E}_h}\Phi_\sigma^e=\Big(\sum_{K\ni\sigma}\sum_{\sigma'\in K}c_{\sigma\sigma'}^K+\sum_{\sigma'\in K}c_{\sigma\sigma'}^K\Big)u_\sigma$$
$$+\sum_{\sigma'}\Bigg(\sum_{K,\sigma,\sigma'\in K}c_{\sigma\sigma'}^K\Bigg)u_{\sigma'}$$

Here, in order to simplify the notations, we have set $c_{\sigma,\sigma'}^{\mathcal{K}}=0$ when $\sigma\notin\mathcal{K}$ or $\sigma'\notin\mathcal{K}$.

The results holds true because $c_{\sigma\sigma'}^{\mathcal{K}}\geq 0$, and

$$\sum_{K\ni\sigma}\sum_{\sigma'\in\mathcal{K}}c_{\sigma\sigma'}^{\mathcal{K}}+\sum_{\sigma'\in\mathcal{K}}c_{\sigma\sigma'}^{\mathcal{K}}=\sum_{\sigma'}\Bigg(\sum_{\mathcal{K},\sigma,\sigma'\in\mathcal{K}}c_{\sigma\sigma'}^{\mathcal{K}}\Bigg).\qquad\square$$

The idea is to construct schemes that satisfy the requirement $c_{\sigma,\sigma'}^{\mathcal{K}}\geq 0$. It is known since Godunov that one cannot have a scheme that is both monotonicity preserving and high order accurate, hence some sort of non linearity must be introduced. Before showing how we can meet the requirements, let us introduce our reference monotone scheme. It is a multidimensional extension of the Rusanov (or local Lax-Friedrichs) scheme, namely, for any $\mathcal{K}$ and σ,

$$\Phi_\sigma^{\mathcal{K}}=\frac{1}{n_{\mathcal{K}}}\Phi^{\mathcal{K}}+\alpha_k\big(u_\sigma-\overline{u}_{\mathcal{K}}\big),\qquad \overline{u}_{\mathcal{K}}=\frac{1}{n_{\mathcal{K}}}\sum_{\sigma\in\mathcal{K}}u_\sigma \tag{19}$$

This scheme has the form (18) and is monotone if $\alpha_K\geq\max_{\mathcal{K}}||\nabla_u\mathbf{f}(u^h)||$.

Another example of monotone residual is called the N scheme (N stands for narrow), and it is due to P.L. Roe in the $\mathbb{P}^1$ case. The construction is as follows. We notice that the total residual on $\mathcal{K}$, thanks to the Gauss formula, also writes

$$\Phi^{\mathcal{K}}=\int_{\mathcal{K}}\text{div }\mathbf{f}(u^h)=\int_{\mathcal{K}}\nabla\mathbf{f}_u(u^h)\cdot\nabla u^h=\sum_{\sigma\in\mathcal{K}}\Bigg(\int_{\mathcal{K}}\nabla\mathbf{f}_u(u^h)\cdot\nabla\varphi_\sigma)\Bigg)u_\sigma$$

We introduce the "inflow" parameters $k_\sigma=\int_{\mathcal{K}}\nabla\mathbf{f}_u(u^h)\cdot\nabla\varphi_\sigma)$, so that $\Phi^{\mathcal{K}}=\sum_\sigma k_\sigma u_\sigma$. We notice that $\sum_\sigma k_\sigma=0$. This parameters are called the inflow parameters because in the $\mathbb{P}^1$ case and for a linear flux, their sign characterizes whether the flow $\nabla_u\mathbf{f}(u^h)$ is inflow or outflow in the element $\mathcal{K}$. The N-scheme is then defined by

$$\Phi_\sigma^N=\max(k_\sigma,0)\big(u_\sigma-\overline{u}\big) \tag{20a}$$

$$\overline{u}=N\Bigg(\sum_{\sigma\in\mathcal{K}}\min(k_\sigma,0)u_\sigma\Bigg) \tag{20b}$$

$$N^{-1}=\sum_{\sigma\in\mathcal{K}}\min(k_\sigma,0) \tag{20c}$$

The average $\overline{u}$ is defined such that the relations (8) hold true. An easy calculation shows that

$$c_{\sigma'\sigma}^N=\min(k_\sigma,0)N\max(k_\sigma,0)\geq 0$$

so that the scheme is monotonicity preserving. Numerical experiments shows that this a very good first order for $\mathbb{P}^1$ element (hence for triangles and tetrahedrons) and provides less good results for higher elements.

Similarly, one can define an upwind high order scheme, nicknamed as the LDA scheme (Low Diffusion A schemes, there has been a LDB, less successful), it is defined by:

$$\Phi_\sigma^{LDA} = -\max(k_\sigma, 0)N\Phi.$$

It is a very good scheme for triangular/tet $\mathbb{P}^1$ elements, but it reveals to be unstable for higher elements or non triangular elements.

3.1. Explicit construction. The construction is local to an element (or boundary edge) $\mathcal{K}$, so we drop the dependency with respect to the element. We start from a monotone first order scheme, such as the Rusanov or the N scheme, denote the first order residuals in the element as $\{\Phi_\sigma^M\}_{\sigma\in K}$ and the high order residuals (to be constructed) by $\{\Phi_\sigma^H\}_\sigma$. We then make the following formal observation:

$$\text{for all } \sigma \in \mathcal{K}, \Phi_\sigma^H = \frac{\Phi_\sigma^H}{\Phi_\sigma^M}\Phi_\sigma^M,$$

so that if $\Phi_\sigma^M = \sum_{\sigma'\in\mathcal{K}} c_{\sigma\sigma'}^M (u_{\sigma'} - u_\sigma)$, we have

$$\begin{aligned}
\phi_\sigma^H &= \frac{\Phi_\sigma^H}{\Phi_\sigma^M}\Big(\sum_{\sigma'\in\mathcal{K}} c_{\sigma\sigma'}^M (u_{\sigma'} - u_\sigma)\Big)\\
&= \sum_{\sigma'\in\mathcal{K}} \left(\frac{\Phi_\sigma^H}{\Phi_\sigma^M} c_{\sigma'\sigma}^M\right)(u_{\sigma'} - u_\sigma)\Big)\\
&= \sum_{\sigma'\in\mathcal{K}} c_{\sigma'\sigma}^H (u_{\sigma'} - u_\sigma)\Big)
\end{aligned}$$

with $c_{\sigma'\sigma}^H := \frac{\Phi_\sigma^H}{\Phi_\sigma^M} c_{\sigma'\sigma}^M$. Hence, to have $c_{\sigma'\sigma}^H \geq 0$, it is enough that

$$\Phi_\sigma^H \ \Phi_\sigma^M \geq 0$$

Introducing the parameters $\beta_\sigma^M = \frac{\Phi_\sigma^M}{\Phi}$ and $\beta_\sigma^H = \frac{\Phi_\sigma^H}{\Phi}$ where Φ is the total residual on the element $\mathcal{K}$, we see that:

- $\Phi_\sigma^H \ \Phi_\sigma^M \geq 0$ is equivalent to $\beta_\sigma^M \ \beta_\sigma^H \geq 0$,
- the conservation relations translates into:

$$\sum_{\sigma\in\mathcal{K}} \beta_\sigma^M = \sum_{\sigma\in\mathcal{K}} \beta_\sigma^H = 1. \tag{21}$$

- In order to guaranty the condition (17), a sufficient condition is that : for any C, and u^h such that $||u^h||_\infty \leq C$, there exists C' such that $|\beta_\sigma^H| \leq C'(C)$, uniformly for all meshes $\mathcal{T}_h$.

These constraints can easily be interpreted geometrically. Consider an simplex $\mathcal{S} = (\mathbf{a}_1, \dots \mathbf{a}_{n_\mathcal{K}})$ of dimension $n_\mathcal{K} - 1$ points, i.e. a triangle when $n_\mathcal{K} = 3$, a tetrahedron for $n_\mathcal{K} = 4$ and so on. These points have nothing to do with the mesh, they are only used to represent easily the constraint (21): it is well known that any point $\mathbf{M}$ of an affine space of dimension $n_{\mathcal{K}}-1$ can be uniquely described n in term of its barycentric coordinates with respect to $\mathcal{S}$:

$$M=\sum_{i=1}^{n_{\mathcal{K}}-1}\lambda_i\mathbf{a}_i, \qquad \sum_{i=1}^{n_{\mathcal{K}}-1}\lambda_i=1$$

so this suggest to interpret the parameters β_σ^M and β_σ^H as barycentric coordinates with respect to the simplex $\mathcal{S}$: we interpret a scheme as a point in this abstract affine space, and finding the mapping $(\beta_\sigma^M)_{\sigma\in\mathcal{K}}\mapsto(\beta_\sigma^H)_{\sigma\in\mathcal{K}}$ can be interpreted as to find a mapping from this affine space onto itself. Then, to make the discussion more visual, we switch to $n_{\mathcal{K}}=3$, see figure 1. The conditions $\beta_\sigma^H\beta_\sigma^L\geq 0$ are interpreted as saying that β_i^H and β_i^L must be on the same side of the line $\lambda_i=0$.

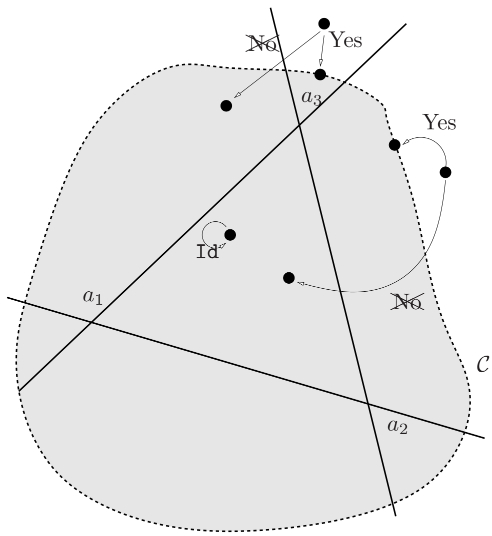

Figure 1. Geometrical representation of the monotonicity conditions. The invariant domain is materialized by the domain inside of $\mathcal{C}$.

The condition $|\beta_\sigma|\leq C$ is materialized, on figure (1), by the domain inside curve $\mathcal{C}$. Inside the invariant domain bounded by $\mathcal{C}$, the mapping is the identity, outside of $\mathcal{C}$ project the point $L=\sum_\sigma\beta_\sigma^L\mathbf{a}_\sigma$ on $\mathcal{C}$ without crossing the lines $\lambda_{\sigma_i}=0$. Once the β_σ^H are defined, we set simply $\Phi_\sigma^H=\beta_\sigma^H\Phi$.

The simplest invariant domain is certainly the simplex $(\mathbf{a}_1,\dots,\mathbf{a}_{n_{\mathcal{K}}})$ for which $0\leq\lambda_\sigma\leq 1$. In that case, the most common formula is [6, 19]:

$$\beta_\sigma^H=\frac{\max(\beta_\sigma^M,0)}{\sum_{\sigma\in\mathcal{K}}\max(\beta_\sigma^M,0)}. \tag{22}$$

Note that $\sum_{\sigma\in\mathcal{K}}\max(\beta_\sigma^M,0)\geq 1$ because

$$1=\sum_{\sigma\in\mathcal{K}}\beta_\sigma^M=\sum_{\sigma\in\mathcal{K}}\max(\beta_\sigma^M,0)+\sum_{\sigma\in\mathcal{K}}\min(\beta_\sigma^M,0)\leq\sum_{\sigma\in\mathcal{K}}\max(\beta_\sigma^M,0).$$

When $\Phi = 0$, we simply set $\Phi_\sigma^H = 0$

In practice, this method is excellent for computing discontinuous solutions. When computing smoother solutions, we can see "wiggles" appearing, see section 5. They are not a manifestation of any instability since the scheme is perfectly L^∞ stable, but it is too over compressive, i.e. not dissipative enough.

It is quite easy to understand what is going on. We first, let us consider the problem on $[0,1]^2$:

$$\frac{\partial u}{\partial x} = 0 \tag{23}$$

with the boundary condition $u = g$ on $\{0\} \times [0,1]$. The grid is made of quadrangles, with vertexes (x_i, y_j), $x_i = \frac{i}{N}$, $y_j = \frac{j}{N}$, $0 \leq i, j \leq N$. The function g is piecewise linear, and $g(0, y_j) = (-1)^j$. The exact solution is independent of x.

The scheme is defined by

$$u_{ij}^{n+1} = u_{ij}^n - \omega_{ij} \sum_{K \ni (x_i, y_j)} \Phi_{i,j}^{H,K}(u_h^n)$$

with u_{ij}^0 given, and $u_{0j}^n = g(0, y_j)$. There are many ways of initializing, we consider two initializations:

- Initialization with the exact solution: $u_{ij}^0 = g(0, y_j) = (-1)^j$
- Check-board mode: $u_{ij}^0 = (-1)^{i+j}$

The solution at the n-th iteration is reconstructed with the $\mathbb{Q}^1$ interpolation. It is easy to see that for both initialization, we have, for any K,

$$\Phi^K = \int_{\partial K} u^h \mathbf{n}_x = 0$$

so that in both cases, for any i, j, n, $u_{ij}^n = u_{ij}^0$! The method, as it is, is not well posed, and there are spurious modes.

To remedy to this serious drawback, there are several possibilities, see [2]. The most flexible one is to add a streamline diffusion term:

$$\Phi_\sigma^{H,K,\star} = \Phi_\sigma^{H,K} + \theta_K h_K \int_K \big(\nabla_u \mathbf{f}(u^h) \cdot \nabla \varphi_\sigma\big)\, N \,\big(\nabla_u \mathbf{f}(u^h) \cdot \nabla u^h\big) \tag{24}$$

where N is define by (20b), and $\theta_K \approx 0$ in discontinuities and $\theta_K \approx 1$ away from discontinuities. When we apply this correction (with $\theta = 1$) to (23) this corrects the problem.

To see what is the rational behind (24), let us first switch to the one dimensional problem:

$$\begin{aligned} \frac{\partial f(u)}{\partial x} &= 0 \quad x \in [0,1] \\ u(0) &= u_0 \\ u(1) &= u_1. \end{aligned} \tag{25}$$

The boundary conditions are imposed weakly, and to make things simple, assume $f'(u_0) > 0$ and $f'(u_1) < 0$ so that the solution is $u = u_0$. The interval $[0,1]$ is discretized with the mesh which elements are $[x_i, x_{i+1}]$, $0 = x_0 < x_1 < \ldots < x_{n-1}, x_n = 1$. Whatever the order, the total residual is for $K_{i+1/2} = [x_i, x_{i+1}]$

$$\Phi^{K_i+1/2} = f(u_{i+1}) - f(u_i)$$

so that the high order residuals are simply, for any degree of freedom $\sigma \in K$, $\Phi_\sigma^K = \beta_\sigma^K\big(f(u_{i+1} - f(u_i)\big)$. In particular, the internal degrees of freedom play no role. Assume now that $k = 1$, there is no internal degree of freedom, and let us evaluate the entropy balance for the entropy $U(u) = \frac{1}{2}u^2$:

$$\begin{aligned}\mathcal{E} &= \sum_{i=0}^{N-1} u_i\bigg(\beta_i^{K_{i-1/2}}\big(f(u_{i+1}) - f(u_i)\big) + \beta_i^{K_{i+1/2}}\big(f(u_{i+1}) - f(u_i)\big)\bigg)\\ &= \int_0^1 u^h\frac{\partial f}{\partial x}(u^h) + \sum_{i=0}^{N-1}\bigg(\gamma_i^{K_{i+1/2}}u_i + \gamma_{i+1}^{K_{i+1/2}}u_{i+1/2}\bigg)\big(f(u_{i+1}) - f(u_i)\big)\end{aligned}$$

with $\gamma_j^{K_{i+1/2}} = \beta_j^{K_{i+1/2}} - \frac{1}{2}$

$$= \int_0^1 u^h\frac{\partial f}{\partial x}(u^h) + \sum_{i=0}^{N-1}\gamma_{i+1}^{K_{i+1/2}}(f(u_{i+1}) - f(u_i))(u_{i+1} - u_i).$$

For the scheme to be dissipative, a sufficient condition is that for all i,

$$\gamma_{i+1}^{K_{i+1/2}}(f(u_{i+1}) - f(u_i))(u_{i+1} - u_i) \geq 0,$$

i.e.

$$\gamma_{i+1}^{K_{i+1/2}}\frac{f(u_{i+1}) - f(u_i)}{u_{i+1} - u_i} \geq 0$$

with a strict inequality for at least one interval.

The evaluation of $\beta_\sigma^{K_{i+1/2}}$ is done with the only aim of having an L^∞ stable scheme, so that this inequality might not be true [1]. Adding the streamline term, i.e.

$$\theta(u_{i+1} - u_i)\int_{x_i}^{x_{i+1}} N\big(\frac{\partial f}{\partial u}\big)^2\frac{\partial\varphi_\sigma}{\partial x} = (u_{i+1} - u_i)\big|\frac{\partial f}{\partial u}\big|\big(\varphi_\sigma(x_{i+1}) - \varphi_\sigma(x_i)\big)$$

will modify the entropy into

$$\mathcal{E} = \int_0^1 u^h\frac{\partial f}{\partial x}(u^h) + \sum_{i=0}^{N-1}\bigg(\gamma_{i+1}^{K_{i+1/2}}\frac{f(u_{i+1}) - f(u_i)}{u_{i+1} - u_i} + \theta\big|\frac{\partial f}{\partial u}\big|\bigg)(u_{i+1} - u_i)^2$$

and $\mathcal{E} \leq \int_0^1 u^h\dfrac{\partial f}{\partial x}(u^h)$ provided that $\theta \geq 1$.

In the general case, we have the following result:

Proposition 3.2. *There exists $\theta > 0$ which depends only on the polynomial degree r such that if $\hat{f}_{\mathbf{n}}$ is an E-flux, then* (12) *is true with the residuals defined by* (24)

[1] However, in 1D it is very simple to show that the sign condition is true, let us ignore this fact however.

Proof. We need to check (12). On the elements K, we get:

$$\begin{aligned}
\sum_{\sigma\in K} u_\sigma\Big(\Phi_\sigma^{H,K} + \theta_K h_K \int_K \big(\nabla \mathbf{f}(u^h)\nabla u^h\big) N\big(\nabla \mathbf{f}(u^h)\nabla u^h\big)\Big)\\
&= \int_{\partial K} \mathbf{g_n} + \sum_\sigma \gamma_\sigma^K (u_\sigma - u_{\sigma_1}) \int_K \text{div } \mathbf{f}(u^h)\\
&\quad + \theta_K h_K \int_K \big(\nabla \mathbf{f}(u^h)\nabla u^h\big) N\big(\nabla \mathbf{f}(u^h)\nabla u^h\big)\Big)\\
&= \int_{\partial K} \mathbf{g_n} + \Big(\sum_{\sigma\in K} \gamma_\sigma^K (u_\sigma - u_{\sigma_1})\Big) \int_K \nabla_u \mathbf{f}(u^h)\cdot \nabla u^h\\
&\quad + \theta h_K \int_K \big(\nabla \mathbf{f}(u^h)\cdot u^h\big)^2 N.
\end{aligned} \tag{26}$$

We see that the second term of the last line can be written as :

$$\Big(u_{\sigma_2} - u_{\sigma_1}, \dots, u_{\sigma_{n_K}} - u_{\sigma_1}\Big)\big(M + \theta_K Q\big)\begin{pmatrix} u_{\sigma_2} - u_{\sigma_1}\\ \vdots \\ u_{\sigma_{n_K}} - u_{\sigma_1}\end{pmatrix}$$

with

$$\mathcal{E}_K = M_{\sigma\sigma'} = \gamma_\sigma^K \int_K \nabla_u f(u^h)\cdot \nabla\varphi_{\sigma'}$$

and

$$Q_{\sigma\sigma'} = h_K \int_K (\nabla_u \mathbf{f}(u^h)\nabla\varphi_\sigma) N (\nabla_u \mathbf{f}(u^h)\nabla\varphi_{\sigma'}).$$

The matrix Q is positive, $\ker Q \subset \ker M$. Since N is constant, we see that

$$\begin{aligned}
&\Big(u_{\sigma_2} - u_{\sigma_1}, \dots, u_{\sigma_{n_K}} - u_{\sigma_1}\Big) M \begin{pmatrix} u_{\sigma_1} - u_{\sigma_1}\\ \vdots \\ u_{\sigma_{n_K}} - u_{\sigma_1}\end{pmatrix}\\
&\geq -\sqrt{|K|} h_K \sqrt{\sum_\sigma (\gamma_\sigma^K)^2}\max_K ||\nabla u^h|| \sqrt{\int_K (\nabla_u \mathbf{f}(u^h)\cdot\nabla u^h)^2}.
\end{aligned}$$

Since $\mathbb{P}^r(K)$ is finite dimensional, there exists $C_{2,\infty}$ which depends only on r such that

$$\sqrt{|K|}\max_K ||\nabla u^h|| \leq C_{2,\infty}\sqrt{\int_K (\nabla u^h)^2}$$

so that

$$\begin{aligned}
\mathcal{E}_K \geq &- h_K C_{2,\infty}\sqrt{\sum_\sigma (\gamma_\sigma^K)^2}\sqrt{\int_K (\nabla u^h)^2}\sqrt{\int_K (\nabla_u \mathbf{f}(u^h)\cdot\nabla u^h)^2}\\
&+ h_K \theta N \int_K (\nabla_u \mathbf{f}(u^h)\cdot \nabla u^h)^2.
\end{aligned}$$

The last thing to show is the existence of $C_r > 0$ such that $\sqrt{\int_K (\nabla_u \mathbf{f}(u^h) \cdot \nabla u^h)^2} \geq C\sqrt{\int_K (\nabla u^h)^2}$ on $\mathbb{P}_r(K)$. Since $\mathbb{P}_r(K) = \ker Q \bigoplus H$ where the two spaces are orthogonal with respect to the scalar product[2] $a(u,v) = \int_K \nabla u \cdot \nabla v$, and because the space are finite dimensional, there exists $C > 0$ such that

$$\forall u \in U_h, \frac{\int_K (\nabla_u \mathbf{f}(u^h) \cdot \nabla u^h)^2}{\int_K (\nabla u^h)^2} \geq C_r > 0.$$

Connecting all the pieces together, since $\beta_\sigma^K \in [0,1]$. $\sum\sum_\sigma (\gamma_\sigma^K)^2 \leq n_K$, we see that $\theta \geq \frac{C_r}{C_{2,\infty}}$ guaranties the entropy inequality.

On the boundary element, if one takes and E-flux, the inequality is also valid. □

Remark 3.3. In practical simulations, $\theta = \frac{1}{n_K}$ is fine.

4. A variational formulation for RD schemes

Though described only by discrete formula, it is possible to identify the mapping χ in (5). Using the same technique, we see that

$$\begin{aligned}\Phi_\sigma^K =& \beta_\sigma^K \int_K \text{div } \mathbf{f}(u^h) + \theta h_K \int_K \big(\nabla_u \mathbf{f}(u^h) \nabla \varphi_\sigma\big) N \big(\nabla_u \mathbf{f}(u^h) \nabla u^h\big) \\ =& \int_K \chi_{u^h}(\varphi_\sigma) \text{ div } \mathbf{f}(u^h),\end{aligned}$$

with

$$\chi_{u^h}(v^h) = \sum_{\sigma \in K} \Bigg(\beta_\sigma^K v_\sigma + \theta h_K \big(\nabla_u \mathbf{f}(u^h) \nabla \varphi_\sigma\big) N \Bigg). \tag{27}$$

5. Numerical examples

In this section, we illustrate the behavior of the method on two examples: a linear transport problem and a non linear one. In $\Omega = [0,1]^2$, we consider

$$\vec{\lambda} = (y, -x)^T \text{ and } u(x,y) = \varphi_0(x) \text{ if } y = 0 \tag{28}$$

with the boundary conditions

$$\varphi_0(x) = \begin{cases} \cos^2(2\pi x) & \text{if } x \in [\frac{1}{4}, \frac{3}{4}] \\ 0 & \text{else} \end{cases}$$

The isolines of the exact solution are circles of center $(0,0)$. The form of the Burgers equation is the following:

$$\begin{aligned} \frac{\partial u}{\partial y} + \frac{1}{2}\frac{\partial u^2}{\partial x} = 0 \qquad & \text{if } x \in [0,1]^2 \\ u(x,y) = 1.5 - 2x \quad & \text{on the inflow boundary.} \end{aligned} \tag{29a}$$

[2]if we remove the subspace of constant polynomial, which is included in $\ker Q$, this becomes a scalar product, thus sum is direct and H depends intrinsically on $\ker Q$.

The exact solution consists in a fan that merges into a shock which foot is located at $(x, y) = (3/4, 1/2)$. More precisely, the exact solution is

$$u(x,y) = \begin{cases} \text{if } y \geq 0.5 & \begin{cases} -0.5 & \text{if } -2(x-3/4)+(y-1/2) \geq 0 \\ 1.5 & \text{else} \end{cases} \\ \text{else} & \max\left(-0.5, \min\left(1.5, \dfrac{x-3/4}{y-1/2}\right)\right) \end{cases} \tag{29b}$$

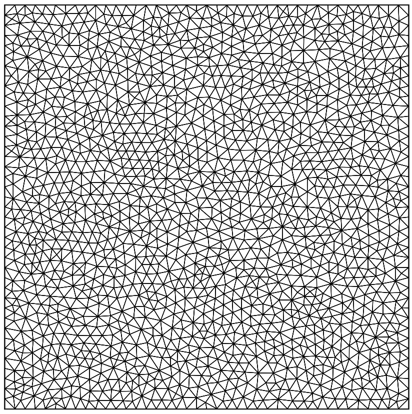

Figure 2. Mesh for the numerical experiments.

The mesh displayed on figure 2 is used to obtain the solutions shown on figure 3 and 4.

We see, on figure 3-(a) that without the streamline term in (24), the solution looks very wiggly. Again, it is not an instability, only a manifestation of spurious modes that are completely eliminated using (24). If one makes a convergence study on this problem using $\mathbb{P}^1$, $\mathbb{P}^2$ and $\mathbb{P}^3$ elements, we recover the expected order of convergence.

h	$\epsilon_{L^2}(\mathbb{P}^1)$	$\epsilon_{L^2}(\mathbb{P}^2)$	$\epsilon_{L^2}(\mathbb{P}^3)$
1/25	0.50493E-02	0.32612E-04	0.12071E-05
1/50	0.14684E-02	0.48741E-05	0.90642E-07
1/75	0.74684E-03	0.13334E-05	0.16245E-07
1/100	0.41019E-03	0.66019E-06	0.53860E-08
	$\mathcal{O}^{\text{ls}}_{L^2}$ =1.790	$\mathcal{O}^{\text{ls}}_{L^2}$ =2.848	$\mathcal{O}^{\text{ls}}_{L^2}$ =3.920

Table 1. Order of accuracy on refined mesh constructed from the mesh of figure 2, L^2 norm. The slopes are obtained by least square

Strictly speaking, the streamline in (24) destroys the maximum preserving nature of the scheme: the operator defined by (24) is not, a priori, of the type (18) with positive coefficients. We have not been able, so far, to analyze in full detail the scheme from this point of view, but all the numerical experiments that we have done so far, including with system case, indicate that the streamline term (24) acts as a filter, and does not spoil the monotonicity preserving properties. Actually, this property is violated, but the over- and undershoot are negligible, as what occurs for the ENO and WENO schemes.

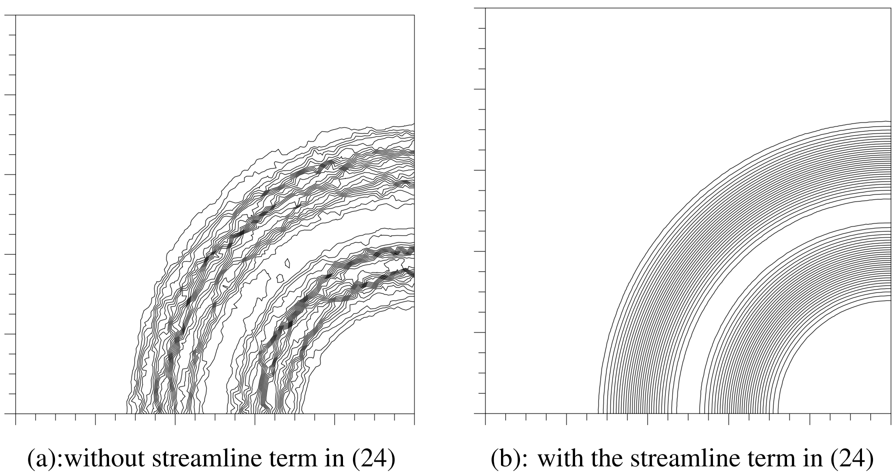

Figure 3. Solution of (28) with (22) and (24), $\mathbb{P}^2$ elements

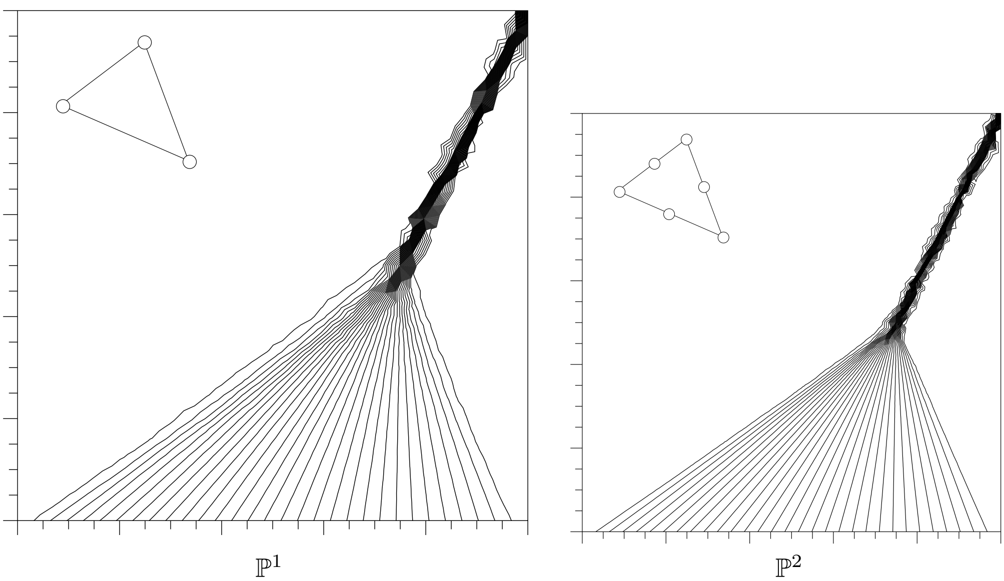

Figure 4. Solution of (29) with (24)

6. Flux formulation of Residual Distribution schemes

In this section we show that the scheme (7) also admits a flux formulation, with an explicit form of the flux. Hence the method is also locally conservative. This is well known for the Finite Volume and Discontinuous Galerkin approximation, much less understood for the RDS and continuous finite elements, despite the paper [12].

Let us consider any common edge or face Γ of K^+ and K^-, two elements. Let $\mathbf{n}$ be the normal to Γ, see Figure 5. A flux $\hat{\mathbf{f}}(S^+, S^-, \mathbf{n})$ between K^+ and K^- has to satisfy

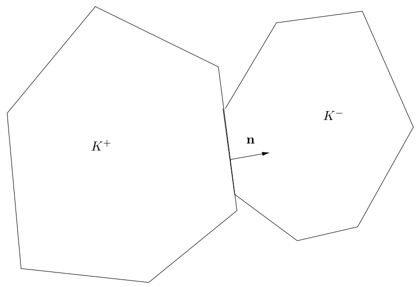

Figure 5.

$$F(S^+, S^-, \mathbf{n}) = -F(S^-, S^+, \mathbf{n}). \tag{30a}$$

and the consistency condition

$$F(S, S, \mathbf{n}) = f(S) \cdot \mathbf{n}. \tag{30b}$$

In (30a), the symbols $S^{\pm}$ represent set of states, where S^+ is associated to K^+ and S^- to K^-. For a first order finite volume scheme, we have $S^+ = \mathbf{u}_{K_+}$ and $S^- = \mathbf{u}_{K_-}$, the average values of $\mathbf{u}$ in K^+ and K^-. For the other schemes the definition is more involved. The aim of this section is to define $\hat{\mathbf{f}}$ and $S^{\pm}$ in the RDS case.

We briefly recall finite volume schemes. Then we show that RDS can be interpreted as finite volume schemes. To make the exposure easier, we assume that $d = 2$ and that the tessellation is conformal, made of triangles. This is not essential as the analysis shows it.

6.1. Analysis.

6.1.1. A recap on Finite volume methods. We denote the list of edges/faces of the elements of $\mathcal{K}$ by $\mathcal{G}$. Considering a numerical flux $\hat{\mathbf{f}}$, and a cell K, the formulation is

$$\int_{\partial K} \mathbf{f}(\mathbf{u}) \cdot \mathbf{n} \approx \sum_{\Gamma \in \mathcal{G}} \hat{\mathbf{f}}(\mathbf{u}^+, \mathbf{u}^-, \mathbf{n}_\Gamma)$$

so that an approximation of (1) is

$$|K| \frac{u_K^{n+1} - u_K^n}{\Delta t} + \sum_{\Gamma \in \mathcal{G}} F(\mathbf{u}^+, \mathbf{u}^-, \mathbf{n}) + \sum_{\Gamma \in \mathcal{G}, \Gamma \subset \partial\Omega^+} F(\mathbf{u}^+, \mathbf{g}, \mathbf{n}) = 0 \tag{31a}$$

and initial conditions

$$u_K^0 = \frac{\int_K u_0(x)dx}{|K|}. \tag{31b}$$

In (31a), we have specialized for the MUSCL method however this is not essential. We have chosen a simple Euler forward time stepping, more accurate solutions can be obtained using the method of lines, for example by using SSP Runge Kutta approximations [10]. More details can be found in [9, 15].

6.1.2. Finite volume as Residual distribution schemes. Here, we rephrase [1]. The notations are defined in Figure 6. Again, we specialize ourself to the case of triangular elements,

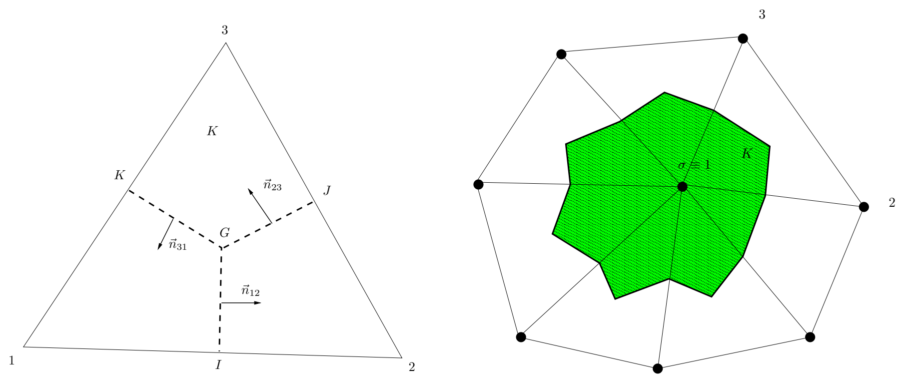

Figure 6. Notations for the finite volume schemes. On the left: definition of the control volume for the degree of freedom σ. The vertex σ plays the role of the vertex 1 on the left picture, etc for the triangle K.

but clearly *exactly the same arguments* can be given for more general elements, provided a conformal approximation space of the type U_h can be constructed. This is clearly the case for triangle elements, and we can take $p = 1$.

The control volume in this case are defined as the median cell, see figure 6. We concentrate on the div $\mathbf{f}$ approximation. Since the boundary of C is a closed polygon, we have

$$\sum_{\gamma \subset \partial C} \mathbf{n}_\gamma = 0$$

where γ is any of the segment included in ∂C, such as IG on Figure 6. Hence

$$\begin{aligned}\sum_{\gamma \subset \partial C} \hat{\mathbf{f}}(u_\sigma^+ u^-, \mathbf{n}_\gamma) &= \sum_{\gamma \subset \partial C} \hat{\mathbf{f}}(u_\sigma^+ u^-, \mathbf{n}_\gamma) - \left(\sum_{\gamma \subset \partial C} \mathbf{n}_\gamma\right) \cdot \mathbf{f}(\mathbf{u}_h(\sigma)) \\ &= \sum_{K, \sigma \in K} \sum_{\text{internal boundaries around } \sigma} \big(\hat{\mathbf{f}}(u_\sigma^+ u^-, \mathbf{n}_\gamma) - \mathbf{f}(\mathbf{u}_h(\sigma)) \cdot \mathbf{n}_\gamma\big)\end{aligned}$$

To make things explicit, in K, the internal boundaries are IG, JG and KG, and those around $\sigma \equiv 1$ are IG and KG. We set

$$\Phi_\sigma^K = \sum_{\text{internal boundaries around } \sigma} \big(\hat{\mathbf{f}}(u_\sigma^+ u^-, \mathbf{n}_\gamma) - \mathbf{f}(\mathbf{u}_h(\sigma)) \cdot \mathbf{n}_\gamma\big). \tag{32}$$

If now we sum up these three quantities and get:

$$
\begin{aligned}
\sum_{\sigma\in K}\Phi_\sigma^K &= \Big(\hat{\mathbf{f}}(u_1^+,u_2^+,\mathbf{n}_{12})-\hat{\mathbf{f}}(u_1^+,u_3^+,\mathbf{n}_{13})-\mathbf{f}(\mathbf{u}_1)\cdot\mathbf{n}_{12}+\mathbf{f}(\mathbf{u}_1)\cdot\mathbf{n}_{31}\Big)\\
&+\Big(\hat{\mathbf{f}}(u_2^+,u_3^+,\mathbf{n}_{23})-\hat{\mathbf{f}}(u_2^+,u_1^+,\mathbf{n}_{12})+\mathbf{f}(\mathbf{u}_2)\cdot\mathbf{n}_{12}-\mathbf{f}(\mathbf{u}_2)\cdot\mathbf{n}_{23}\Big)\\
&+\Big(-\hat{\mathbf{f}}(u_3^+,u_2^+,\mathbf{n}_{23})+\hat{\mathbf{f}}(u_3^+,u_1^+,\mathbf{n}_{31})-\mathbf{f}(\mathbf{u}_3)\cdot\mathbf{n}_{23}+\mathbf{f}(\mathbf{u}_3)\cdot\mathbf{n}_{31}\Big)\\
&=\mathbf{f}(\mathbf{u}_1)\cdot\big(\mathbf{n}_{12}-\mathbf{n}_{31}\big)+\mathbf{f}(\mathbf{u}_2)\cdot\big(-\mathbf{n}_{23}+\mathbf{n}_{31}\big)+\mathbf{f}(\mathbf{u}_3)\cdot\big(\mathbf{n}_{31}-\mathbf{n}_{23}\big)\\
&=\mathbf{f}(\mathbf{u}_1)\cdot\frac{\mathbf{n}_1}{2}+\mathbf{f}(\mathbf{u}_2)\cdot\frac{\mathbf{n}_2}{2}+\mathbf{f}(\mathbf{u}_3)\cdot\frac{\mathbf{n}_3}{2}
\end{aligned}
$$

where $\mathbf{n}_j$ is the scaled inward normal of the edge opposite to vertex σ_j, i.e. twice the gradient of the $\mathbb{P}^1$ basis function φ_{σ_j} associated to this degree of freedom. Thus, we can reinterpret the sum as the boundary integral of the Lagrange interpolant of the flux. The finite volume scheme is then a residual distribution scheme with residual defined by (32) and a total residual defined by

$$
\Phi^K := \int_{\partial K}\mathbf{f}^h\cdot\mathbf{n},\qquad \mathbf{f}^h=\sum_{\sigma\in K}\mathbf{f}(\mathbf{u}_\sigma)\varphi_\sigma. \tag{33}
$$

6.1.3. Residual distribution schemes as finite volume schemes.. Let K be a fixed triangle. We are given a set of residues $\{\Phi_\sigma^K\}_{\sigma\in K}$, our aim here is to define a flux function such that relations similar to (32) hold true. We show the method for $\mathbb{P}^1$ and $\mathbb{P}^2$ interpolant, more general cases can easily be handled in the same way.

Warm up: The $\mathbb{P}^1$ case. Let us begin with the $\mathbb{P}^1$ case: the degrees of freedom are the vertexes of K, and we consider a linear interpolation in K. The flux across ID in the direction $\mathbf{n}_{12}$ is denoted by $\hat{\mathbf{f}}_{\mathbf{n}_{12}}$ and the flux across IG in the direction $-\mathbf{n}_{12}$ is $\hat{\mathbf{f}}_{-\mathbf{n}_{12}}=-\hat{\mathbf{f}}_{\mathbf{n}_{12}}$ by definition. Using similar notations, we must satisfy

$$
\begin{aligned}
\Phi_1 &= \hat{\mathbf{f}}_{\mathbf{n}_{12}}-\hat{\mathbf{f}}_{\mathbf{n}_{31}}-\mathbf{f}(\mathbf{u}_1)\cdot\frac{\mathbf{n}_1}{2}\\
\Phi_2 &= -\hat{\mathbf{f}}_{\mathbf{n}_{12}}+\hat{\mathbf{f}}_{\mathbf{n}_{23}}-\mathbf{f}(\mathbf{u}_2)\cdot\frac{\mathbf{n}_2}{2}\\
\Phi_3 &= -\hat{\mathbf{f}}_{\mathbf{n}_{23}}+\hat{\mathbf{f}}_{\mathbf{n}_{31}}-\mathbf{f}(\mathbf{u}_3)\cdot\frac{\mathbf{n}_3}{2}
\end{aligned} \tag{34}
$$

Clearly, there is a compatibility relation:

$$
\Phi^K=\sum_\sigma \mathbf{f}(\mathbf{u}_\sigma)\cdot\nabla\varphi_\sigma. \tag{35}
$$

We can rewrite (34) as a linear system:

$$
\begin{pmatrix}\Phi_1+\mathbf{f}(\mathbf{u}_1)\cdot\frac{\mathbf{n}_1}{2}\\ \Phi_2+\mathbf{f}(\mathbf{u}_2)\cdot\frac{\mathbf{n}_2}{2}\\ \Phi_3+\mathbf{f}(\mathbf{u}_3)\cdot\frac{\mathbf{n}_3}{2}\end{pmatrix}
=\begin{pmatrix}1&-1&0\\-1&0&1\\0&1&-1\end{pmatrix}
\begin{pmatrix}\hat{\mathbf{f}}_{\mathbf{n}_{12}}\\ \hat{\mathbf{f}}_{\mathbf{n}_{31}}\\ \hat{\mathbf{f}}_{\mathbf{n}_{23}}\end{pmatrix}
:= A\begin{pmatrix}\hat{\mathbf{f}}_{\mathbf{n}_{12}}\\ \hat{\mathbf{f}}_{\mathbf{n}_{31}}\\ \hat{\mathbf{f}}_{\mathbf{n}_{23}}\end{pmatrix}
$$

The matrix A is not invertible but has rank 2. Since (35) is true, there exists one solution at least. We can find easily one sample solution.

Let us first set $\hat{\mathbf{f}}_{\mathbf{n}_{31}} = 0$. Then we get

$$\begin{aligned}
\hat{\mathbf{f}}_{\mathbf{n}_{12}} &= \Phi_1 + \mathbf{f}(\mathbf{u}_1) \cdot \tfrac{\mathbf{n}_1}{2} \\
\hat{\mathbf{f}}_{\mathbf{n}_{23}} &= \Phi_1 + \Phi_2 + \mathbf{f}(\mathbf{u}_2) \cdot \tfrac{\mathbf{n}_2}{2} + \mathbf{f}(\mathbf{u}_1) \cdot \tfrac{\mathbf{n}_1}{2} \\
\hat{\mathbf{f}}_{\mathbf{n}_{31}} &= 0
\end{aligned}$$

Thanks to (35), this can be rewritten as

$$\begin{aligned}
\hat{\mathbf{f}}_{\mathbf{n}_{12}} &= \Phi_1 + \mathbf{f}(\mathbf{u}_1) \cdot \tfrac{\mathbf{n}_1}{2} \\
\hat{\mathbf{f}}_{\mathbf{n}_{23}} &= -\Phi_3 - \mathbf{f}(\mathbf{u}_3) \cdot \tfrac{\mathbf{n}_3}{2} \\
\hat{\mathbf{f}}_{\mathbf{n}_{31}} &= 0
\end{aligned}$$

Then we set $\hat{\mathbf{f}}_{\mathbf{n}_{12}} = 0$, thus

$$\begin{aligned}
\hat{\mathbf{f}}_{\mathbf{n}_{12}} &= 0 \\
\hat{\mathbf{f}}_{\mathbf{n}_{23}} &= \Phi_2 + \mathbf{f}(\mathbf{u}_2) \cdot \tfrac{\mathbf{n}_2}{2} \\
\hat{\mathbf{f}}_{\mathbf{n}_{31}} &= -\Phi_1 + \mathbf{f}(\mathbf{u}_1) \cdot \tfrac{\mathbf{n}_1}{2}.
\end{aligned}$$

Last, we set $\hat{\mathbf{f}}_{\mathbf{n}_{23}} = 0$ and get

$$\begin{aligned}
\hat{\mathbf{f}}_{\mathbf{n}_{12}} &= -\Phi_2 - \mathbf{f}(\mathbf{u}_2) \cdot \tfrac{\mathbf{n}_2}{2} \\
\hat{\mathbf{f}}_{\mathbf{n}_{23}} &= 0 \\
\hat{\mathbf{f}}_{\mathbf{n}_{31}} &= \Phi_3 + \mathbf{f}(\mathbf{u}_3) \cdot \tfrac{\mathbf{n}_3}{2}
\end{aligned}$$

To have a symmetric formulation, it is enough to take the average,

$$\begin{aligned}
\hat{\mathbf{f}}_{\mathbf{n}_{12}} &= \frac{\Phi_1 - \Phi_2}{3} + \frac{1}{6}\Big(\mathbf{f}(\mathbf{u}_1) \cdot \mathbf{n}_1 - \mathbf{f}(\mathbf{u}_2) \cdot \mathbf{n}_2\Big) \\
\hat{\mathbf{f}}_{\mathbf{n}_{23}} &= -\frac{\Phi_2 - \Phi_3}{3} + \frac{1}{6}\Big(\mathbf{f}(\mathbf{u}_2) \cdot \mathbf{n}_2 - \mathbf{f}(\mathbf{u}_3) \cdot \mathbf{n}_3\Big) \\
\hat{\mathbf{f}}_{\mathbf{n}_{31}} &= \frac{\Phi_3 - \Phi_1}{3} + \frac{1}{6}\Big(\mathbf{f}(\mathbf{u}_3) \cdot \mathbf{n}_3 - \mathbf{f}(\mathbf{u}_1) \cdot \mathbf{n}_2\Big)
\end{aligned}$$

or, by introducing $\Psi_i = \Phi_i - \mathbf{f}(\mathbf{u}_i) \cdot \frac{\mathbf{n}_i}{2}$,

$$\hat{\mathbf{f}}_{\mathbf{n}_{12}} = \frac{1}{3}\big(\Psi_1 - \Psi_2\big), \quad \hat{\mathbf{f}}_{\mathbf{n}_{23}} = \frac{1}{3}\big(\Psi_2 - \Psi_3\big), \quad \hat{\mathbf{f}}_{\mathbf{n}_{31}} = \frac{1}{3}\big(\Psi_3 - \Psi_1\big). \tag{36}$$

Let us check the consistency of the flux. We first have to adapt the notion of consistency. As recalled in the Introduction, two of the key arguments in the proof of the Lax-Wendroff theorem are related to the structure of the flux, for classical finite volume schemes. In [6], the proof is adapted to the case of Residual Distribution schemes. The property that stands for the consistency is that if in an element, all the states are identical, then the residuals are all vanishing. Hence, we will say that

Definition 6.1. A multidimensional flux

$$\hat{\mathbf{f}} := \hat{\mathbf{f}}(\mathbf{u}_1, \ldots, \mathbf{u}_n, \mathbf{n})$$

is consistent if, when $\mathbf{u}_1 = \mathbf{u}_2 = \ldots = \mathbf{u}_n = \mathbf{u}$ then

$$\hat{\mathbf{f}}(\mathbf{u}, \ldots, \mathbf{u}, \mathbf{n}) = \mathbf{f}(\mathbf{u}) \cdot \mathbf{n}.$$

Let us show that the flux (36) are consistent in that sense. If the three states are equal to $\mathbf{u}$, then we have

$$\hat{\mathbf{f}}_{\mathbf{n}_{12}} = \frac{1}{6}\mathbf{f}(\mathbf{u}) \cdot \Big(\mathbf{n}_1 - \mathbf{n}_2\Big), \quad \hat{\mathbf{f}}_{\mathbf{n}_{23}} = \frac{1}{6}\mathbf{f}(\mathbf{u}) \cdot \Big(\mathbf{n}_2 + \mathbf{n}_3\Big), \quad \hat{\mathbf{f}}_{\mathbf{n}_{31}} = \frac{1}{6}\mathbf{f}(\mathbf{u}) \cdot \Big(\mathbf{n}_3 - \mathbf{n}_2\Big)$$

By symmetry, we only consider the first relation. Using the notations of the figure 6, we see that $\mathbf{n}_1 - \mathbf{n}_2$ is the normal of $\vec{BC} - \vec{CA} = \vec{BC} + \vec{AC}$. Since G is the centroid of the triangle, we see that $\vec{GC} = (\vec{AC} + \vec{BC})/3$, and thus we get

$$\hat{\mathbf{f}}_{\mathbf{n}_{12}} = \mathbf{f}(\mathbf{u}) \cdot \mathbf{n}_{12}.$$

This ends the proof.

We can state a couple of general remarks:

1. In general, the residuals depends on more than 2 arguments. For stabilized finite element methods, or the non linear stable residual distribution schemes, see e.g. [5, 12, 19], the residuals depends on the three states of K. Thus the formula (36) shows that the flux on more than two states in contrast to the 1D case. In the Finite volume case however, the support of the flux function is generally larger than the three states of K, think for example of an ENO/WENO method, of a simpler MUSCL one.
2. The formula (36) are influenced by the form of the total residual (33). We show in the next paragraph how this can be generalized.
3. We have set at the beginning that $\hat{\mathbf{f}}_{\mathbf{n}_{ij}} = -\hat{\mathbf{f}}_{-\mathbf{n}_{ij}}$. The formula (36) are antisymmetric with respect to the indices, and then do respect the assumed equality.

The example of the $\mathbb{P}^2$ approximation and the more general case. We consider the set-up defined by Figure 7. The triangle is splitted first into 4 sub-triangles K_1, K_2, K_3 and K_4. From this sub-triangulation, we can construct a dual mesh as in the $\mathbb{P}^1$ case and we have represented the 6 sub-zones that are the intersection of the dual control volumes and the triangle K. Our notations are as follow: given any sub-triangle K_ξ, if γ_{ij} is intersection between two adjacent control volumes (associated to σ_i and σ_j vertices of K_ξ), the normal to γ_{ij} in the direction σ_i to σ_j is denoted by $\mathbf{n}_{ij}^\xi$. Similarly the flux across γ_{ij} is denoted $\hat{\mathbf{f}}_{ij}^\xi$.

Following the same method as in the $\mathbb{P}^1$ case, we set:

$$\begin{array}{lll}
\Phi_1 & = \hat{\mathbf{f}}_{14}^1 - \hat{\mathbf{f}}_{61}^1 & + \int_{\partial C_1 \cap K} \mathbf{f}(\mathbf{u}^h) \cdot \mathbf{n} \\
\Phi_2 & = -\hat{\mathbf{f}}_{42}^2 + \hat{\mathbf{f}}_{25}^2 & + \int_{\partial C_2 \cap K} \mathbf{f}(\mathbf{u}^h) \cdot \mathbf{n} \\
\Phi_3 & = -\hat{\mathbf{f}}_{53}^3 + \hat{\mathbf{f}}_{36}^3 & + \int_{\partial C_3 \cap K} \mathbf{f}(\mathbf{u}^h) \cdot \mathbf{n} \\
\Phi_4 & = -\hat{\mathbf{f}}_{14}^1 + \big(\hat{\mathbf{f}}_{46}^1 - \hat{\mathbf{f}}_{64}^4\big) + \big(\hat{\mathbf{f}}_{45}^4 - \hat{\mathbf{f}}_{54}^2\big) + \hat{\mathbf{f}}_{42}^2 & + \int_{\partial C_4 \cap K} \mathbf{f}(\mathbf{u}^h) \cdot \mathbf{n} \\
\Phi_5 & = -\hat{\mathbf{f}}_{25}^2 + \big(\hat{\mathbf{f}}_{54}^2 - \hat{\mathbf{f}}_{45}^4\big) + \big(\hat{\mathbf{f}}_{56}^4 - \hat{\mathbf{f}}_{65}^3\big) + \hat{\mathbf{f}}_{53}^3 & + \int_{\partial C_5 \cap K} \mathbf{f}(\mathbf{u}^h) \cdot \mathbf{n} \\
\Phi_6 & = -\hat{\mathbf{f}}_{36}^3 + \big(\hat{\mathbf{f}}_{65}^3 - \hat{\mathbf{f}}_{56}^4\big) + \big(\hat{\mathbf{f}}_{64}^4 - \hat{\mathbf{f}}_{46}^1\big) + \hat{\mathbf{f}}_{61}^1 & + \int_{\partial C_6 \cap K} \mathbf{f}(\mathbf{u}^h) \cdot \mathbf{n}.
\end{array} \tag{37}$$

We can group the terms in (37) by sub-triangles, namely:

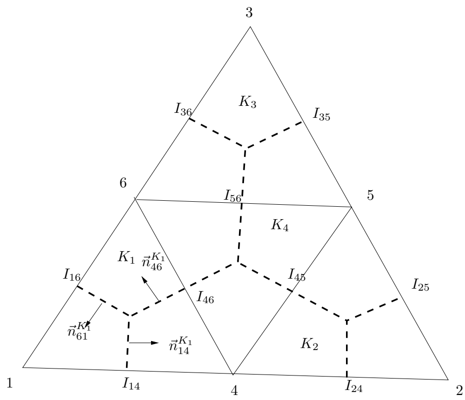

Figure 7. Geometrical elements for the $\mathbb{P}^2$ case.

$$
\begin{aligned}
\Phi_1 =& \left(\hat{\mathbf{f}}^1_{14} - \hat{\mathbf{f}}^1_{61} + \int_{\partial C_1 \cap K_1} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n} \right) \\
\Phi_2 =& \left(-\hat{\mathbf{f}}^2_{42} + \hat{\mathbf{f}}^2_{25} + \int_{\partial C_2 \cap K_2} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n} \right) \\
\Phi_3 =& \left(-\hat{\mathbf{f}}^3_{53} + \hat{\mathbf{f}}^3_{36} + \int_{\partial C_3 \cap K_3} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n} \right) \\
\Phi_4 =& \left(-\hat{\mathbf{f}}^1_{14} + \hat{\mathbf{f}}^1_{46} + \int_{\partial C_4 \cap K_1} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n} \right) \\
&+ \left(-\hat{\mathbf{f}}^4_{64} + \hat{\mathbf{f}}^4_{45} + \int_{\partial C_4 \cap K_4} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n} \right) \\
\Phi_5 =& \left(-\hat{\mathbf{f}}^2_{25} + \hat{\mathbf{f}}^2_{54} + \int_{\partial C_5 \cap K_2} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n} \right) \\
&+ \left(-\hat{\mathbf{f}}^4_{45} + \hat{\mathbf{f}}^4_{56} + \int_{\partial C_5 \cap K_4} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n} \right) \\
&+ \left(-\hat{\mathbf{f}}^3_{65} + \hat{\mathbf{f}}^3_{53} + \int_{\partial C_5 \cap K_3} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n} \right) \\
\Phi_6 =& \left(-\hat{\mathbf{f}}^3_{36} + \hat{\mathbf{f}}^3_{65} + \int_{\partial C_6 \cap K_3} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n} \right) \\
&+ \left(-\hat{\mathbf{f}}^4_{56} + \hat{\mathbf{f}}^4_{64} + \int_{\partial C_6 \cap K_4} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n} \right) \\
&+ \left(-\hat{\mathbf{f}}^1_{46} + \hat{\mathbf{f}}^1_{61} + \int_{\partial C_6 \cap K_1} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n} \right)
\end{aligned}
\tag{38}
$$

where we have used:

$$\begin{aligned}
\int_{\partial C_4\cap K} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n} &= \int_{\partial C_4\cap K_1} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n} + \int_{\partial C_4\cap K_4} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n} + \int_{\partial C_4\cap K_2} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}\\
\int_{\partial C_5\cap K} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n} &= \int_{\partial C_5\cap K_2} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n} + \int_{\partial C_5\cap K_4} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n} + \int_{\partial C_5\cap K_3} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}\\
\int_{\partial C_6\cap K_6} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n} &= \int_{\partial C_6\cap K_3} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n} + \int_{\partial C_6\cap K_4} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n} + \int_{\partial C_6\cap K_1} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}
\end{aligned}$$

Then we define the sub-residuals per sub elements:

$$\begin{aligned}
\Phi_1^1 &= -\hat{\mathbf{f}}_{61}^1 + \hat{\mathbf{f}}_{14}^1 + \int_{\partial C_1\cap K_1} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}, & \Phi_4^2 &= -\hat{\mathbf{f}}_{54}^2 + \hat{\mathbf{f}}_{42}^2 + \int_{\partial C_4\cap K_2} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}\\
\Phi_4^1 &= -\hat{\mathbf{f}}_{14}^1 + \hat{\mathbf{f}}_{46}^1 + \int_{\partial C_4\cap K_1} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}, & \Phi_2^2 &= -\hat{\mathbf{f}}_{42}^2 + \hat{\mathbf{f}}_{25}^2 + \int_{\partial C_2\cap K_2} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}\\
\Phi_6^1 &= -\hat{\mathbf{f}}_{46}^1 + \hat{\mathbf{f}}_{61}^1 + \int_{\partial C_6\cap K_1} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}, & \Phi_5^2 &= -\hat{\mathbf{f}}_{25}^2 + \hat{\mathbf{f}}_{54}^2 + \int_{\partial C_5\cap K_2} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}\\
\\
\Phi_5^3 &= -\hat{\mathbf{f}}_{65}^3 + \hat{\mathbf{f}}_{53}^3 + \int_{\partial C_5\cap K_3} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}, & \Phi_4^4 &= -\hat{\mathbf{f}}_{64}^4 + \hat{\mathbf{f}}_{45}^4 + \int_{\partial C_4\cap K_4} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}\\
\Phi_3^3 &= -\hat{\mathbf{f}}_{36}^3 + \hat{\mathbf{f}}_{65}^3 + \int_{\partial C_6\cap K_3} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}, & \Phi_5^4 &= -\hat{\mathbf{f}}_{45}^4 + \hat{\mathbf{f}}_{56}^4 + \int_{\partial C_5\cap K_4} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}\\
\Phi_6^3 &= -\hat{\mathbf{f}}_{36}^3 + \hat{\mathbf{f}}_{65}^3 + \int_{\partial C_6\cap K_3} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}, & \Phi_6^4 &= -\hat{\mathbf{f}}_{56}^4 + \hat{\mathbf{f}}_{64}^4 + \int_{\partial C_6\cap K_4} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}.
\end{aligned} \tag{39}$$

Clearly,

$$\begin{aligned}
\Phi_1^1 + \Phi_4^1 + \Phi_6^1 &= \int_{\partial K_1} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}, & \Phi_4^2 + \Phi_2^2 + \Phi_5^2 &= \int_{\partial K_2} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}\\
\Phi_5^3 + \Phi_3^3 + \Phi_6^3 &= \int_{\partial K_3} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}, & \Phi_4^4 + \Phi_5^4 + \Phi_6^4 &= \int_{\partial K_4} \mathbf{f}(\mathbf{u}^h)\cdot \mathbf{n}
\end{aligned} \tag{40}$$

so we are back to the $\mathbb{P}^1$ case: in each sub-triangle, we can define flux that will depend on the 6 states of the element via the boundary flux. This is legitimate because in the $\mathbb{P}^1$ case, we have not used the fact that the interpolation is linear, we have only used the fact that we have 3 vertices. Clearly the flux are consistent in the sense of definition 6.1.

The same argument can be clearly extended to higher degree element, as well as to non triangular element: what is needed is to subdivide the element into sub-triangles.

Acknowledgments. The author has been partially supported by the FP7 ERC Advanced Grant ADDECCO # 226316 and the Swiss SNF grant # 200021_153604/1. I would like to thank Mario Ricchiuto, INRIA: our discussions have been very helpfull to simplify some of the arguments developped in this paper.

References

[1] R. Abgrall, *Toward the ultimate conservative scheme: Following the quest,* J. Comput. Phys., **167**(2) (2001), 277–315.

[2] ______, *Essentially non-oscillatory residual distribution schemes for hyperbolic problems,* J. Comput. Phys., **214**(2) (2006), 773–808.

[3] R. Abgrall and D. de Santis, *Linear and non-linear high order accurate residual distribution schemes for the discretization of the steady compressible navier-stokes equations*, J. Comput. Phys., 2014. in revision.

[4] R. Abgrall and D. de Santis, *High-order preserving residual distribution schemes for advection-diffusion scalar problems on arbitrary grids*, SIAM I. Sci. Comput., in press. also http://hal.inria.fr/docs/00/76/11/59/PDF/8157.pdf.

[5] R. Abgrall, A. Larat, and M. Ricchiuto, *Construction of very high order residual distribution schemes for steady inviscid flow problems on hybrid unstructured meshes,* J. Comput. Phys., 230(11) (2011), 4103–4136.

[6] R. Abgrall and P. L. Roe, *High-order fluctuation schemes on triangular meshes,* J. Sci. Comput., **19**(1-3) (2003), 3–36.

[7] B. Cockburn, S. Hou, and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous finite element method for conservation laws IV: the multidimensional case*, Math. Comp., **54** (1990), 545–581.

[8] B. Cockburn and C.-W. Shu, *TVB Runge-Kutta* local projection discontinuous Galerkin finite element method for conservation laws II: General framework. Math. Comp., **52** (1989), 411–435.

[9] E. Godlewski and P.-A. Raviart, *Numerical approximation of hyperbolic systems of conservation laws.* New York, NY: Springer, 1996.

[10] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretisation methods*, SIAM Review, **43**(1) (2001), 89–112.

[11] A. Harten, *On a class of high resolution total-variational-stable finite-difference schemes (with appendix by Peter D. Lax).* SIAM J. Numer. Anal., **21** (1984), 1–23.

[12] T.J.R. Hughes, L.P. Franca, and M. Mallet, *A new finite element formulation for CFD: I*, symmetric forms of the compressible Euler and Navier-Stokes equations and the second law of thermodynamics. Comp. Meth. Appl. Mech. Engrg., **54** (1986), 223–234.

[13] C. Johnson, U. Nävert, and J. Pitkäranta, Finite element methods for linear hyperbolic problems. *Computer methods in applied mechanics and engineering*, **45** (1985), 285–312.

[14] D. Kröner, M. Rokyta, and M. Wierse, *A Lax-Wendroff type theorem for upwind finite volume schemes in* 2-*d*, East-West J. Numer. math., **4**(4) (1996), 279–292.

[15] R. J. Leveque, *Finite volume methods for hyperbolic problems,* Cambridge: Cambridge University Press, 2002.

[16] R.-H. Ni, A multiple grid scheme for solving the Euler equations. In *5th Computational Fluid Dynamics Conference*, pages 257–264, 1981.

[17] P.L. Roe, *Approximate Riemann solvers, parameter vectors, and difference schemes,* J. Comput. Phys., **43** (1981), 357–372.

[18] ______, Characteristic-based schemes for the Euler equations. Annu. Rev. Fluid Mech. **18** (1986), 337–365.

[19] R. Struijs, H. Deconinck, and P.L. Roe, Fluctuation splitting schemes for the 2D Euler equations. VKI-LS 1991-01, 1991. Computational Fluid Dynamics.

Institut für Mathematik, Universität Zürich, Winterthurerstrasse 190, CH-8057 Zürich
E-mail: remi.abgrall@math.uzh.ch